Exploring the Evidence of Speech Recognition and Shorter Passage Length in Computerized Oral Reading Fluency (CORE)

Joseph F. T. Nese

Akihito Kamata

Julie Alonzo

July, 2015

Society for the Scientific Study of Reading

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140203 to the University of Oregon. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

Assessing reading fluency is critical because it functions as an indicator of comprehension and overall reading achievement. Although theory and research demonstrate the importance of ORF proficiency, traditional ORF assessment practices are lacking as sensitive measures of progress for educators to make instructional decisions. The purpose of this study is to compare traditional ORF measures/administration to a computerized ORF assessment system based on speech recognition software (CORE). Using WCPM scores as the outcome, we compare: (a) traditional ORF passages to CORE passages, (b) CORE passage lengths, and (c) scoring methods. We used a general linear model with two within-subject factors, to test the mean WCPM score differences between passage length, scoring method, and their interaction. We found that CORE passages, whether short, medium, or long, tended to have higher WCPM means than the Traditional ORF passages. Real-time scoring tended to have higher WCPM means than both Audio Recording and ASR scoring types, and the ASR and Recording scores were quite similar across passage length and grade, providing preliminary evidence that the speech recognition scoring engine can score passages as well as human administrators in real settings.

Reading fluency, defined as reading a text fluidly, accurately, and with prosody– is a central component in learning to read. Here we focus on accuracy, the extent to which a passage is correctly read, and rate, the time elapsed when reading a passage. Assessing reading fluency is critical because it functions as an indicator of comprehension and overall reading achievement. Measures of ORF are used across the country as part of response to intervention (RTI), a model based on a multi-level prevention framework that involves assessment and intervention to increase student achievement. ORF assessments are used to universally screen for students at risk of low reading proficiency to ensure students are meeting teacher expectations. They are also used to monitor student progress and modify instruction based on data as needed.

There are several practical and methodological limitations of the current and standard ORF assessment. Some of the practical inadequacies of traditional ORF implementation are being discussed in this panel. Also, the opportunity cost of individual administration of traditional ORF assessments is quite high in terms of time and school resources. Depending on the size of the school, screening assessments can take anywhere from one day to an entire week during each the assessment periods (fall, winter, and spring). Instructional time is lost, and there are costs associated with training and paying staff to administer and score ORF assessments.

ORF measures need to be of equivalent difficulty to effectively measure growth. Traditionally, ORF passage difficulty has been determined by estimates of readability, followed up by studies of alternate form reliability and modifications to the passages under development as needed. Dr. Francis, today's Discussant, authored the seminal study on the implications of ORF form effects and passage equivalence. Research places the standard error of ORF measures around 10 WCPM, which yields a 95% confidence interval range of about 40 WCPM. This range is often larger than the magnitude of within-year growth. This is problematic when these measures are used to monitor student progress across time, because true growth is difficult to distinguish from measurement error. Although theory and research demonstrate the importance of ORF proficiency, traditional ORF assessment practices are lacking as sensitive measures of progress for educators to make instructional decisions.

Computerized Oral Reading Evaluation (CORE)

The larger purpose of the Computerized Oral Reading Evaluation (CORE) project is to develop and validate a computerized ORF assessment system to address these practical and methodological limitations. There are two main components: (a) automated scoring based on speech recognition that measures both accuracy and speededness, and (b) a latent variable psychometric model to decrease the standard error of current ORF measures.

We are currently in the first phase of development which includes creating and refining CORE passages, and incorporating automated scoring into the system. We created passages of shorter lengths under the assumption that administering multiple passages can increase the reliability of the score, much like increasing the number of items does.

Purpose

The purpose of this study is to compare traditional ORF measures/administration to a computerized ORF assessment system based on speech recognition software (CORE). Using WCPM scores as the outcome, we compare: (a) traditional ORF passages to CORE passages, (b) CORE passage lengths, and (c) scoring methods.

Method

The ORF passages used in easyCBM were developed to assess students' ability to fluently read narrative text. During instrument development, each form was created to be consistent in length and the readability of each form was verified to fit appropriate grade-level, initially using the Flesch-Kincaid index feature available on Microsoft Word (e.g., Alonzo & Tindal, 2008), with later empirical support through applications in the field. The easyCBM assessment system includes 20 alternate ORF forms of *reported* equivalent difficulty at each grade level, with three forms specifically identified for use as universal screeners (fall, winter, spring) and 17 alternate forms for progress monitoring. During administration, students are given one minute to read as many words as possible in a connected narrative passage (approximately 250 words). These traditional ORF measures have demonstrated features of technical adequacy that suggest they are sufficient to meet the needs of our proposed studies as the comparative example of an existing traditional ORF system (e.g., Alonzo, Mariano, Nese, & Tindal, 2010; Alonzo & Tindal, 2009; Jamgochian et al., 2010; Lai et al., 2010; Saéz et al., 2010).

Students are given one minute to read as many words as possible in a grade-level text while a trained assessor follows along and indicates on a scoring protocol each word the student reads incorrectly (Wayman, Wallace, Wiley, Tichá, & Espin, 2007). If a student pauses for more than three seconds, the assessor prompts the student to continue and marks the word as read incorrectly. Student self-corrections are not marked as errors, but word omissions are. After one minute, the assessor calculates WCPM by subtracting the number of incorrectly read words from the total number of words read.

Each CORE passage is an original work of fiction, ±5 words of the target length. Passages were written to have a beginning, middle, and end in a problem/resolution format, or a sequence of events format. This broad specification is intended to give the passage writers freedom in meeting the word constraint specification, which is crucial in this project. Passages were written with grade-appropriate vocabulary and word frequency so that an average of several well-respected readability scores was estimated to be at grade-level. If passages were found not to be the appropriate level, they were revised to reflect the appropriate level (i.e., increase or decrease level). There were three CORE passage lengths: 9 long passages (about 85 words); 15 medium passages (about 50 words), and 30 short passages (about 25 words).

The CORE system administered, scored, and recorded all passages. There were three scoring methods: (a) Real-time score: a trained assessor scored (WCPM) and the time duration for each passage; (b) Recording score: a trained assessor scored (WCPM) the audio recordings of each passage in a quiet environment, with headphones, and the ability to rewind or confer with another assessor; and (c) Automated Speech Recognition (ASR) score: a speech recognition engine.

Sample

The sample consisted of 342 students in Grades 2 through 4 from two public school districts in the Pacific Northwest. See Table 1.

Table 1

Grade	**Classrooms**	**Students**
2	6	109
3	7	133
4	6	100
Total	**19**	**342**

The traditional ORF passages was always presented first, followed by the CORE passages, grouped by passage length and presented in alternate order. Students were give one minute to read the traditional ORF passage, and all CORE passages were to be read in their entirety. The groups of CORE passage by length (10 short, 5 medium, and 3 long) totaled approximately 250 words, matching the length of the traditional ORF passage. See Table 2.

Table 2

#	Passage Type	Total Words	Time	Order
1	Traditional ORF (easyCBM)	≈ 250	60 seconds	Always first
10	Short CORE	≈ 250	Until completed	Grouped by passage length, alternate order
5	Medium CORE	≈ 250	Until completed	
3	Long CORE	≈ 250	Until completed	

Analyses

We used a general linear model with two within-subject factors, to test the mean WCPM score differences between passage LENGTH, scoring METHOD, and their interaction. The LENGTH factor included 4 categories SHORT, MEDIUM, LONG, AND Traditional ORF. The score METHOD factor included 3 categories: Real-time, Recorded audio, and ASR. Note that we used the Greenhouse--Geisser correction for sphericity, and the Sidak adjustment for the simple paired-effect comparisons to control the familywise error rate.

Results

Please note that results represent preliminary findings.

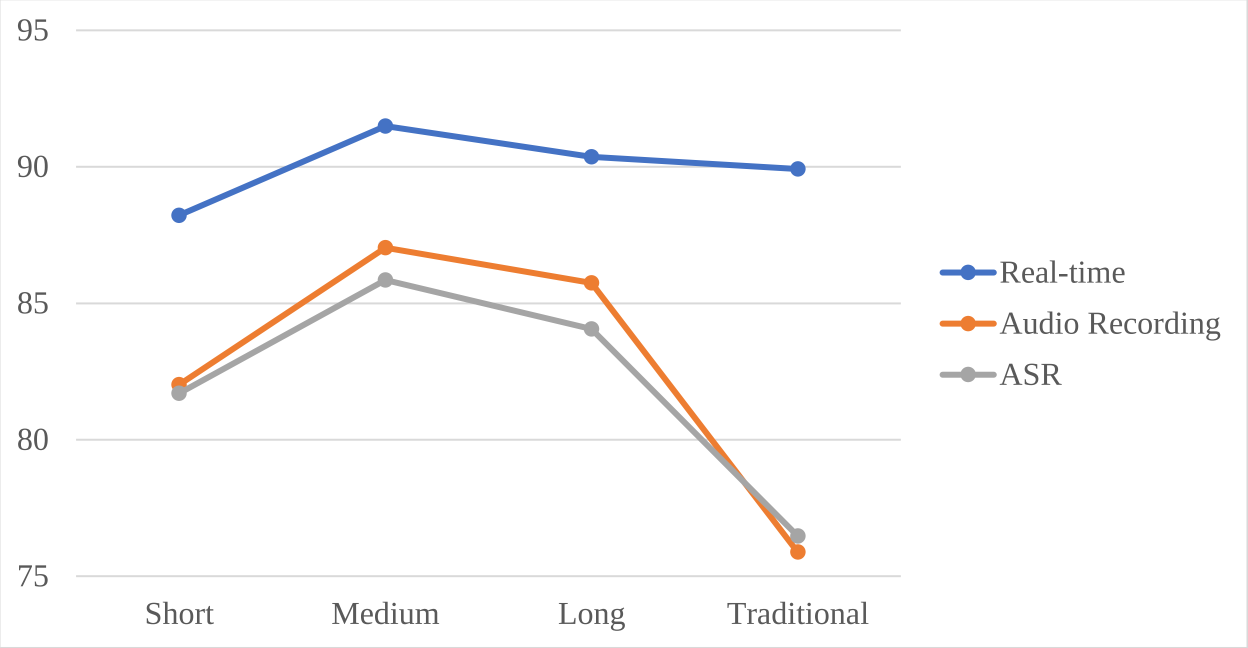
Grade 2. Table 3 shows the table of observed means and standard deviations by scoring method and passage length, and the graph of the estimated marginal means.

*Table 3. Observed Mean WCPM (*SD*)*

	Short	Medium	Long	Traditional
	25 words	50 words	85 words	250 words
Real-time	88.22 (32.75)	91.49 (34.08)	90.37 (35.65)	89.91 (32.02)
Recording	82.02 (34.03)	87.03 (33.36)	85.74 (35.31)	75.88 (29.23)
ASR	81.70 (35.64)	85.85 (32.55)	84.05 (34.20)	76.47 (28.50)

In Grade 2, we found significant main effects for both passage LENGTH and scoring METHOD, and a significant interaction effect. Traditional ORF passage was clearly different from each of the three CORE passage lengths (except under the Real-time scoring method). The Traditional ORF passage mean was lower than the mean all three CORE passage lengths. Among the CORE passages, the "medium" and “long" passages were similar, and the "short" passage was significantly different from the other two. The ASR and Audio Recording scores nearly identical for all passage lengths, and the real-time scores were higher than both the ASR and Recording scores across lengths. Looking at the Simple Effect Comparisons, we found no significant effects for any length comparison under the Real-time scoring method. See Figure 1.

Figure 1. Grade 2 Estimated Marginal Means



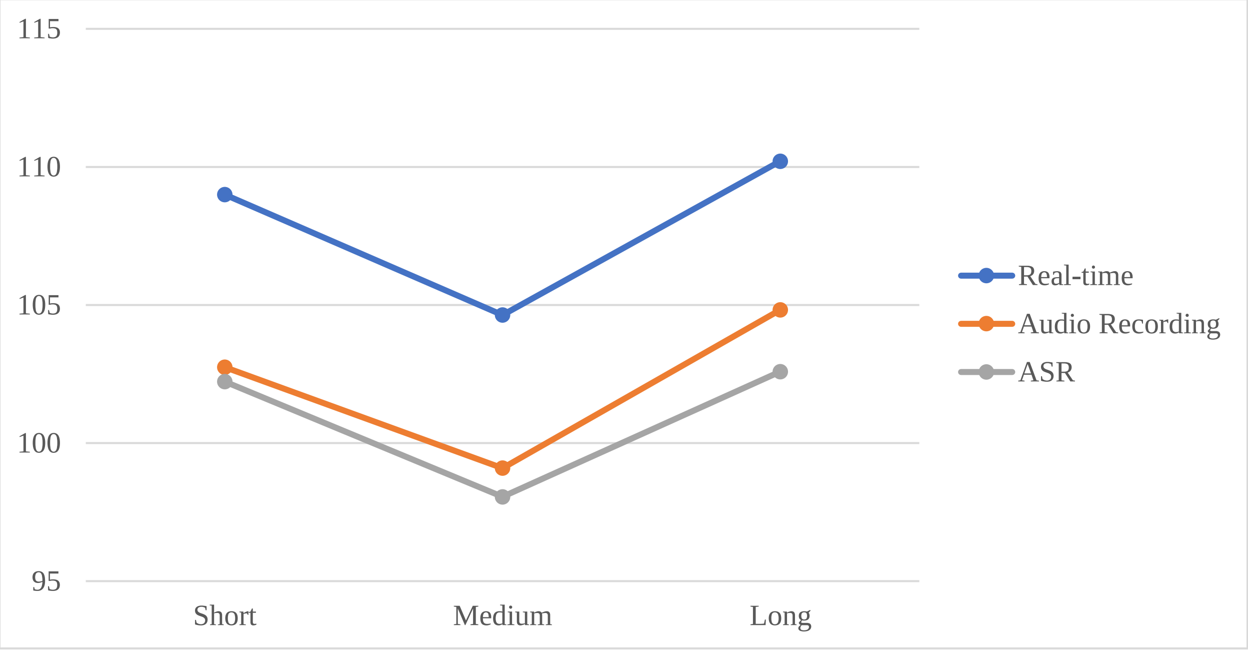
Grade 3. Table 4 shows the table of observed means and standard deviations by scoring method and passage length, and the graph of the estimated marginal means.

*Table 4. Observed Mean WCPM (*SD*)*

	Short 25 words	Medium 50 words	Long 85 words	Traditional 250 words
Real-time	109.00 (34.57)	104.63 (31.91)	110.21 (35.48)	n/a
Recording	102.75 (32.85)	99.09 (30.36)	104.83 (33.59)	n/a
ASR	102.22 (31.63)	98.05 (29.71)	102.59 (32.53)	n/a

In Grade 3 we did not recover scores for the Traditional ORF passages, so here we are only comparing LENGTHS of the CORE passages, SHORT, MEDIUM, and LONG. We found significant main effects for both passage LENGTH and scoring METHOD, and NO significant interaction effect, perhaps because the traditional ORF passages were excluded. Similar to Grade 2, across passage lengths, the ASR and Audio Recording scores nearly identical, while the real-time scores were higher than both the ASR and Recording scores. Among the CORE passages, the “Short" and “Long" lengths were similar, and “Medium" length was different from the other two across all scoring methods.

Figure 2. Grade 3 Estimated Marginal Means



Grade 4. Table 5 shows the table of observed means and standard deviations by scoring method and passage length, and the graph of the estimated marginal means.

Table 5. Observed Mean WCPM (SD)

	Short 25 words	Medium 50 words	Long 85 words	Traditional 250 words
Real-time	134.50 (40.42)	135.38 (41.41)	131.05 (42.11)	122.63 (36.78)
Recording	126.71 (36.89)	127.67 (38.28)	126.26 (40.22)	107.74 (33.01)
ASR	125.62 (35.01)	126.13 (36.85)	125.12 (38.98)	107.18 (32.10)

For Grade 4, we again found significant main effects for both passage LENGTH and scoring METHOD, and a significant interaction effect. Similar to Grade 2, traditional ORF passage means was significantly different, and lower, than each of the three CORE passage lengths. Here, the three CORE passage lengths means were similar to each other across scoring methods. Only for the Real-time scoring method was the LONG passage significantly different from the Medium. And again, the ASR and Audio Recording scores were nearly identical for all passage lengths. And the real-time scores were higher than both the ASR and Recording scores.

Figure 3. Grade 4 Estimated Marginal Means

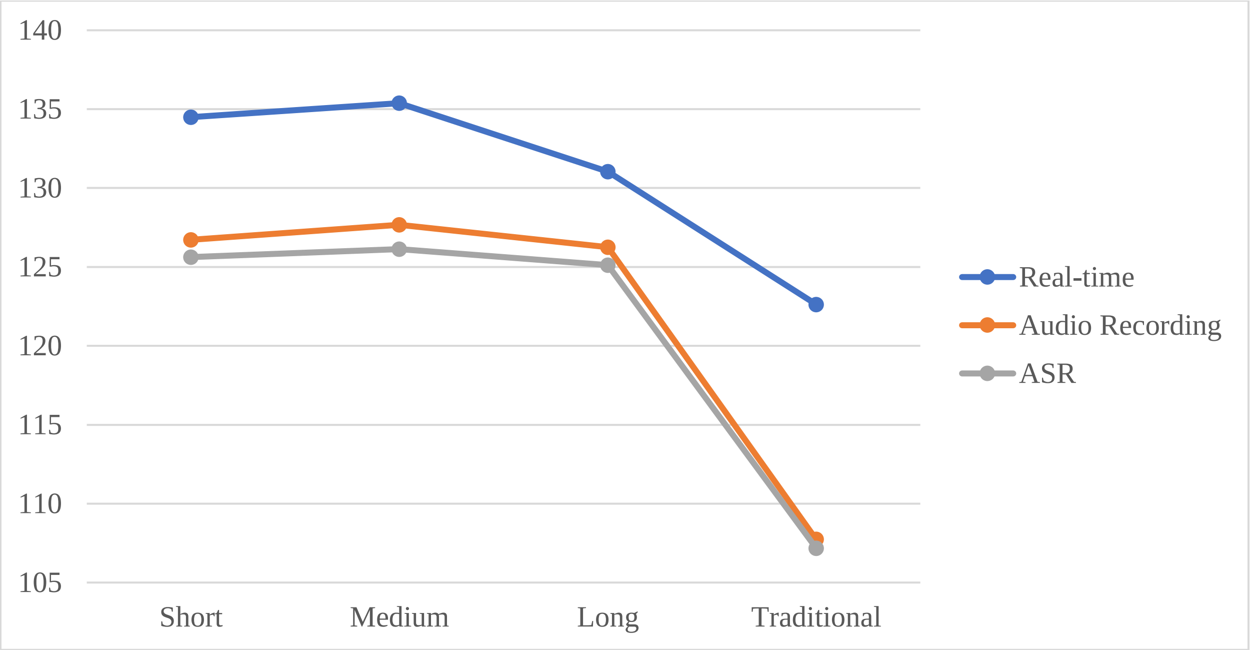


Table 6 shows the correlations across passage lengths for each of the scoring methods, and the ranges are comparable, showing evidence for reliability of scoring across passage length. Thus, in future work on this project, we will make decisions about which passage lengths we will include in the system.

Figure 6. Correlations Across Passage Lengths for the Scoring Methods

Grade	Real-time	Audio Recording	ASR
2	.89-.95	.88-.95	.87-.95
3	.88-.93	.86-.91	.87-.92
4	.90-.96	.84-.95	.86-.94
2-3	.88-.96	.84-.95	.86-.95

Discussion

Overall, it was clear that CORE passages, whether SHORT, MEDIUM, or LONG, tended to have higher WCPM means than the Traditional ORF passages. That is, across grades, students tended to read these passages at a faster WCPM rate than Traditional ORF passages. This was not simply a matter of passage length, as the average WCPM scores did decrease as the passage word count increased. Also, Real-time scoring tended to have higher WCPM means than both Audio Recording and ASR scoring types, and the ASR and Recording scores were quite similar across passage length and grade. This was perhaps the result we were most interested in, to see if the speech recognition engine could accurately score ORF passages. Our intent was to compare the Audio Recording scores to both the Real-time and ASR scores. All else equal, we considered the Audio Recordings to be the “gold standard” score, because scoring could take place in a quiet setting with no distractions, and the capability to rewind the recording to ensure the most accurate word scores. The assumption was that both the Real-time and ASR are susceptible to more errors, so were less concerned with how those compared to each other, than how each compared to the Audio Recording score. The correlation table shows the correlations of the audio recording scores WITH real-time and ASR scores, respectively, and you can see that the correlations are quite similar. This provides some preliminary evidence that the speech recognition scoring engine can score passages as well as human administrators in real settings. So, these results show some promise that the ASR can be used to score ORF passages moving forward.

There are some challenges in scoring ORFs. There are problems with real-time ORF administration, given in noisy environment, where it is hard to hear students, and easy to miss a word. Errors can include: mistakenly scoring a word; timing the reading for more/less than 60 seconds; incorrectly marking the last word read in the allotted time; incorrectly calculating the number of words read correctly; and recording the wrong WCPM score in the database.

But the computerized ORF is only as good as the audio quality. There was no data lost in the real-time scores, however the ASR and audio scores were dependent on the presence of audio, and scores depend on the quality of that audio. Some recordings were lost to poor wifi, or occasional bugs when the audio did not upload. Students would click DONE before reading the entire passage. And although we considered the Audio Recording the "gold standard" score, even with audio, it is sometimes hard to tell when a student reads a word incorrectly, or the student plays with the microphone so quality goes down.

Once we have the complete data from this study, including Grade 3 traditional ORF scores, we will conduct other analyses, like comparing the error rate across passage length and scoring method. We can also do agreement analyses across scoring methods at the word level, compare human and machine timing of passages, and others.

References

Alonzo, J., Mariano, G. J., Nese, J. F., & Tindal, G. (2010). *Reliability of the easyCBM© reading assessments*. Paper presented at the Pacific Coast Research Conference, San Diego, CA.

Alonzo, J., & Tindal, G. (2008). *The development of fifth-grade passage reading fluency measures in a progress monitoring assessment system* (Technical Report No. 43). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2009). *Alternate form and test-retest reliability of easyCBM reading measures* (Technical Report No. 0906). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Jamgochian, E. M., Park, B. J., Nese, J. F. T., Lai, C. F., Saez, L., Anderson, D., Alonzo, J., & Tindal, G. (2010). *Technical Adequacy of the easyCBM Grade 2 Reading Measures* (Technical Report No. 1004). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C.F., Nese, J.F.T., Jamgochian, E.M., Kamata, A., Anderson, D., Park, B.J., Alonzo, J., & Tindal, G. (2010). *Technical adequacy of the easyCBM primary-level reading measures (Grades K-1), 2009-2010 version.* (Technical Report No. 1003). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Saez, L., Park, B. J., Nese, J. F. T., Jamgochian, E. M., Lai, C. F., Anderson, D., Kamata, A., Alonzo, J., & Tindal, G. (2010). *Technical Adequacy of the easyCBM Reading Measures (Grades 3-7), 2009-2010 Version* (Technical Report No. 1005). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41,* 85-120.